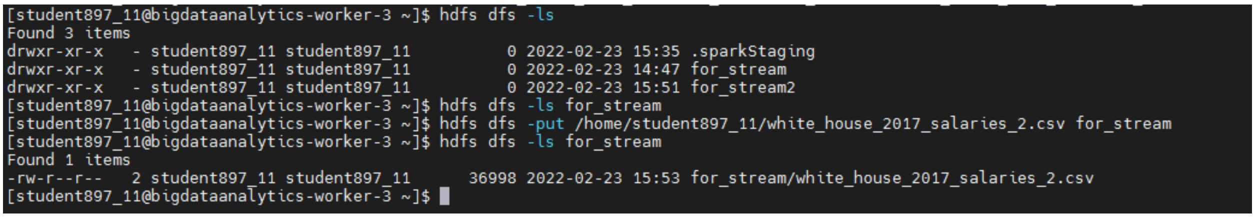
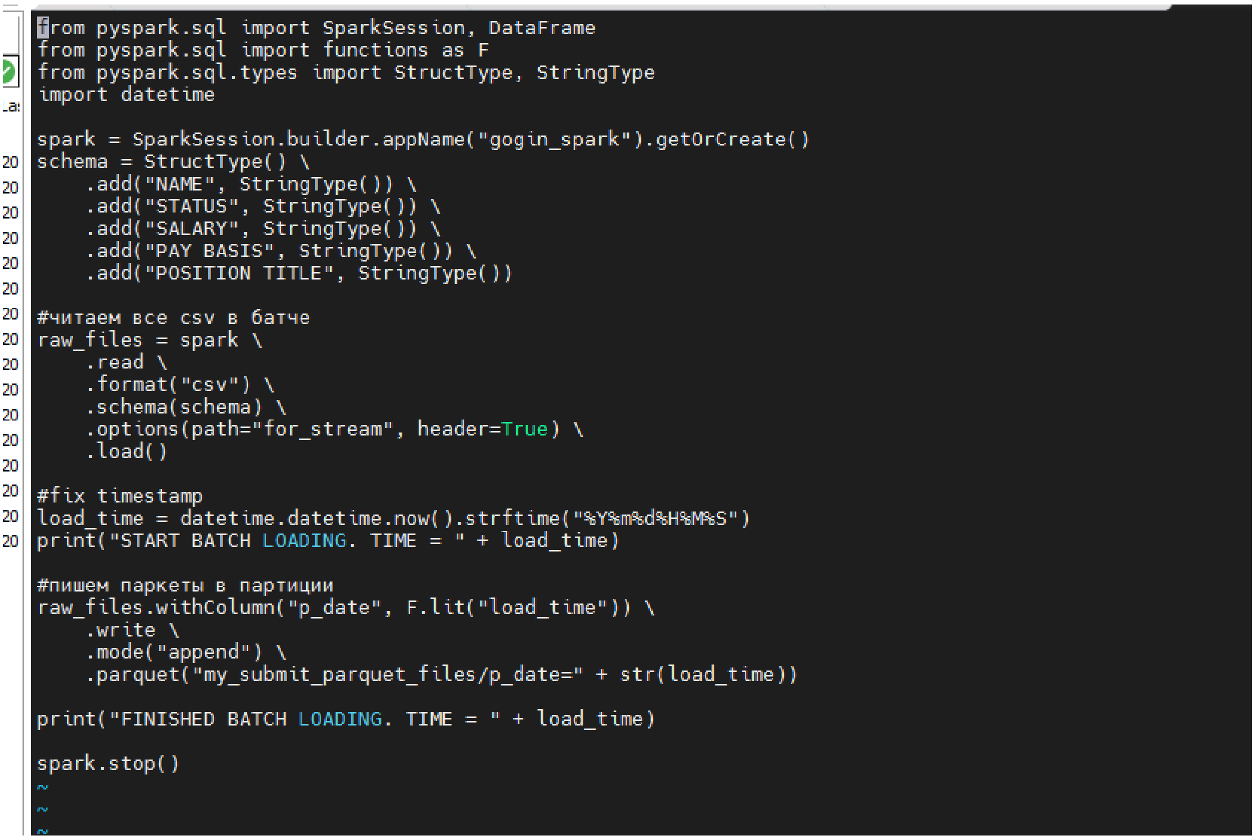
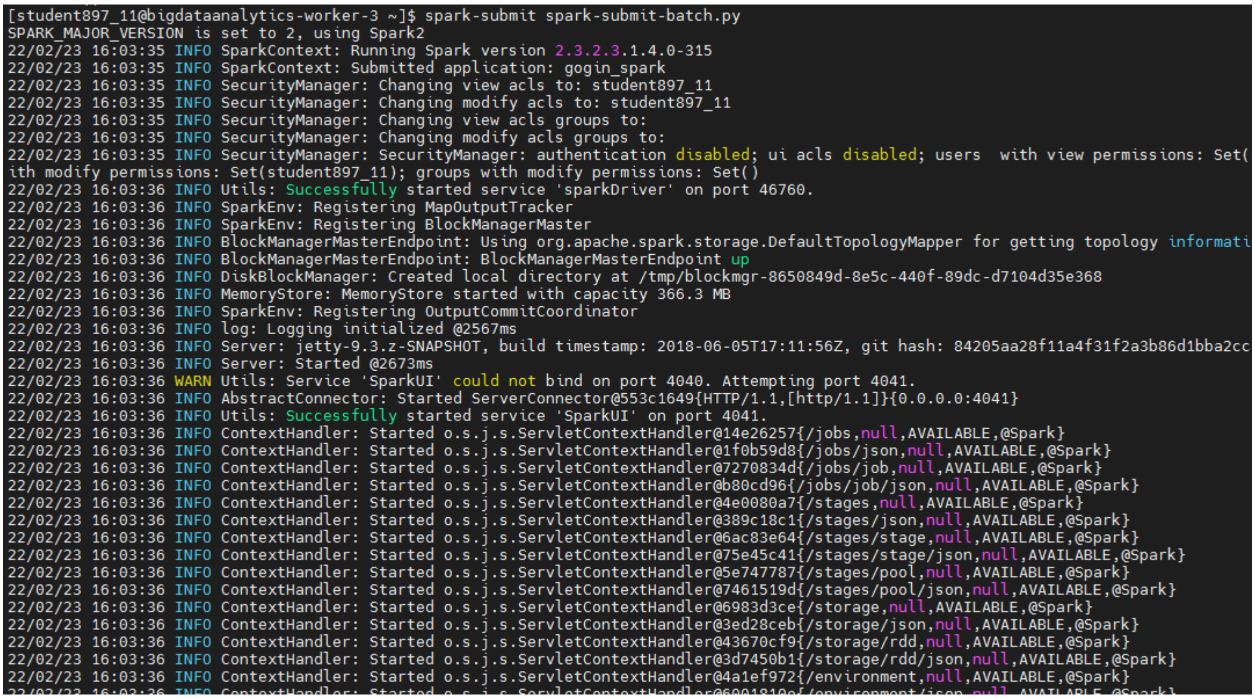


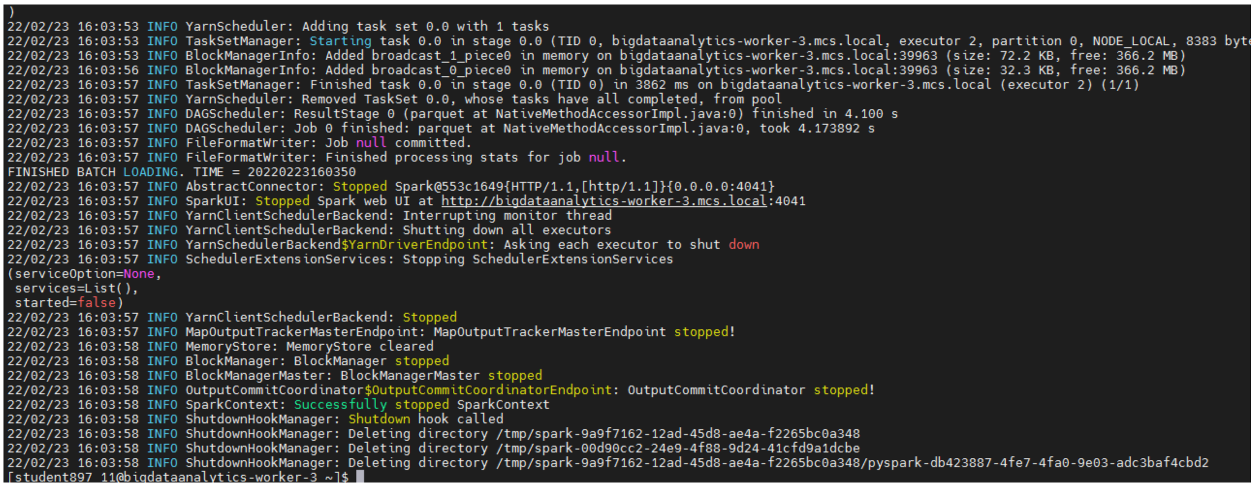
```
[student897_11@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 3 items
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 15:35 .sparkStaging
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 14:47 for_stream
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 15:51 for_stream2
[student897_11@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls for_stream
[student897_11@bigdataanalytics-worker-3 ~]$ hdfs dfs -put /home/student897_11/white_house_2017_salaries_2.csv for_stream
[student897_11@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls for_stream
Found 1 items
-rw-r--r--   2 student897_11 student897_11      36998 2022-02-23 15:53 for_stream/white_house_2017_salaries_2.csv
[student897_11@bigdataanalytics-worker-3 ~]$
```



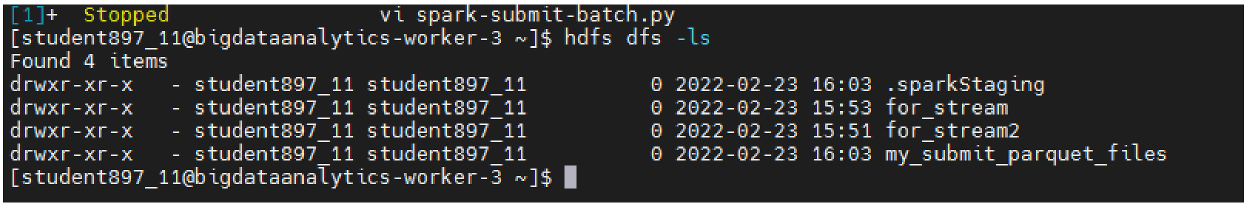
```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

spark = SparkSession.builder.appName("gogin_spark").getOrCreate()
schema = StructType() \
    .add("NAME", StringType()) \
    .add("STATUS", StringType()) \
    .add("SALARY", StringType()) \
    .add("PAY BASIS", StringType()) \
    .add("POSITION TITLE", StringType())

#читаем все csv в батче
raw_files = spark \
    .read \
    .format("csv") \
    .schema(schema) \
    .options(path="for_stream", header=True) \
    .load()

#fix timestamp
load_time = datetime.datetime.now().strftime("%Y%m%d%H%M%S")
print("START BATCH LOADING. TIME = " + load_time)

#пишем паркеты в партиции
raw_files.withColumn("p_date", F.lit("load_time")) \
    .write \
    .mode("append") \
    .parquet("my_submit_parquet_files/p_date=" + str(load_time))

print("FINISHED BATCH LOADING. TIME = " + load_time)

spark.stop()
~
~
~
```



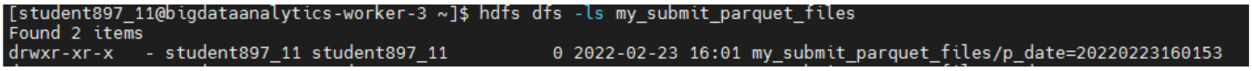
```
[student897_11@bigdataanalytics-worker-3 ~]$ spark-submit spark-submit-batch.py
SPARK_MAJOR_VERSION is set to 2, using Spark2
22/02/23 16:03:35 INFO SparkContext: Running Spark version 2.3.2.3.1.4.0-315
22/02/23 16:03:35 INFO SparkContext: Submitted application: gogin_spark
22/02/23 16:03:35 INFO SecurityManager: Changing view acls to: student897_11
22/02/23 16:03:35 INFO SecurityManager: Changing modify acls to: student897_11
22/02/23 16:03:35 INFO SecurityManager: Changing view acls groups to:
22/02/23 16:03:35 INFO SecurityManager: Changing modify acls groups to:
22/02/23 16:03:35 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(
ith modify permissions: Set(student897_11); groups with modify permissions: Set()
22/02/23 16:03:36 INFO Utils: Successfully started service 'sparkDriver' on port 46760.
22/02/23 16:03:36 INFO SparkEnv: Registering MapOutputTracker
22/02/23 16:03:36 INFO SparkEnv: Registering BlockManagerMaster
22/02/23 16:03:36 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology informati
22/02/23 16:03:36 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/02/23 16:03:36 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-8650849d-8e5c-440f-89dc-d7104d35e368
22/02/23 16:03:36 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
22/02/23 16:03:36 INFO SparkEnv: Registering OutputCommitCoordinator
22/02/23 16:03:36 INFO log: Logging initialized @2567ms
22/02/23 16:03:36 INFO Server: jetty-9.3.z-SNAPSHOT, build timestamp: 2018-06-05T17:11:56Z, git hash: 84205aa28f11a4f31f2a3b86d1bba2cc
22/02/23 16:03:36 INFO Server: Started @2673ms
22/02/23 16:03:36 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
22/02/23 16:03:36 INFO AbstractConnector: Started ServerConnector@553c1649{HTTP/1.1,[http/1.1]}{0.0.0.0:4041}
22/02/23 16:03:36 INFO Utils: Successfully started service 'SparkUI' on port 4041.
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@14e26257{/jobs,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@1f0b59d8{/jobs/json,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@7270834d{/jobs/job,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@b80cd96{/jobs/job/json,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@4e0080a7{/stages,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@389c18c1{/stages/json,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@6ac83e64{/stages/stage,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@75e45c41{/stages/stage/json,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@5e747787{/stages/pool,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@7461519d{/stages/pool/json,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@6983d3ce{/storage,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@3ed28ceb{/storage/json,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@43670cf9{/storage/rdd,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@3d7450b1{/storage/rdd/json,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@4a1ef972{/environment,null,AVAILABLE,@Spark}
22/02/23 16:03:36 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@6001810e{/environment/json,null,AVAILABLE,@Spark}
```



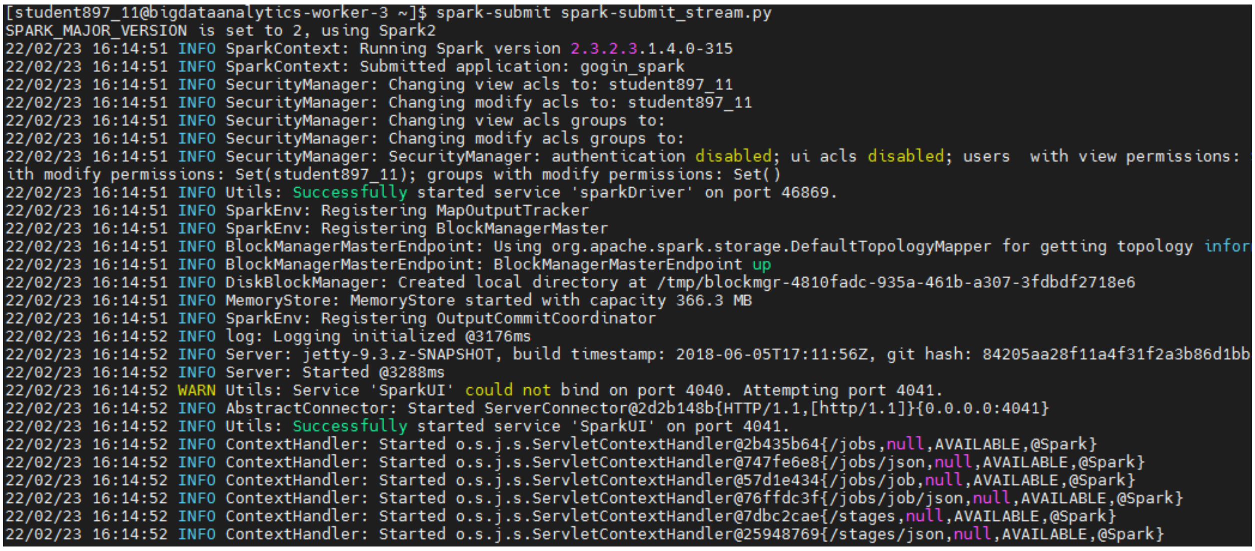
```
)
22/02/23 16:03:53 INFO YarnScheduler: Adding task set 0.0 with 1 tasks
22/02/23 16:03:53 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, bigdataanalytics-worker-3.mcs.local, executor 2, partition 0, NODE_LOCAL, 8383 byte
22/02/23 16:03:53 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on bigdataanalytics-worker-3.mcs.local:39963 (size: 72.2 KB, free: 366.2 MB)
22/02/23 16:03:56 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on bigdataanalytics-worker-3.mcs.local:39963 (size: 32.3 KB, free: 366.2 MB)
22/02/23 16:03:57 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 3862 ms on bigdataanalytics-worker-3.mcs.local (executor 2) (1/1)
22/02/23 16:03:57 INFO YarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool
22/02/23 16:03:57 INFO DAGScheduler: ResultStage 0 (parquet at NativeMethodAccessorImpl.java:0) finished in 4.100 s
22/02/23 16:03:57 INFO DAGScheduler: Job 0 finished: parquet at NativeMethodAccessorImpl.java:0, took 4.173892 s
22/02/23 16:03:57 INFO FileFormatWriter: Job null committed.
22/02/23 16:03:57 INFO FileFormatWriter: Finished processing stats for job null.
FINISHED BATCH LOADING. TIME = 20220223160350
22/02/23 16:03:57 INFO AbstractConnector: Stopped Spark@553c1649{HTTP/1.1,[http/1.1]}{0.0.0.0:4041}
22/02/23 16:03:57 INFO SparkUI: Stopped Spark web UI at http://bigdataanalytics-worker-3.mcs.local:4041
22/02/23 16:03:57 INFO YarnClientSchedulerBackend: Interrupting monitor thread
22/02/23 16:03:57 INFO YarnClientSchedulerBackend: Shutting down all executors
22/02/23 16:03:57 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
22/02/23 16:03:57 INFO SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
22/02/23 16:03:57 INFO YarnClientSchedulerBackend: Stopped
22/02/23 16:03:57 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/02/23 16:03:58 INFO MemoryStore: MemoryStore cleared
22/02/23 16:03:58 INFO BlockManager: BlockManager stopped
22/02/23 16:03:58 INFO BlockManagerMaster: BlockManagerMaster stopped
22/02/23 16:03:58 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/02/23 16:03:58 INFO SparkContext: Successfully stopped SparkContext
22/02/23 16:03:58 INFO ShutdownHookManager: Shutdown hook called
22/02/23 16:03:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-9a9f7162-12ad-45d8-ae4a-f2265bc0a348
22/02/23 16:03:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-00d90cc2-24e9-4f88-9d24-41cfd9a1dcbe
22/02/23 16:03:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-9a9f7162-12ad-45d8-ae4a-f2265bc0a348/pyspark-db423887-4fe7-4fa0-9e03-adc3baf4cbd2
[student897_11@bigdataanalytics-worker-3 ~]$
```



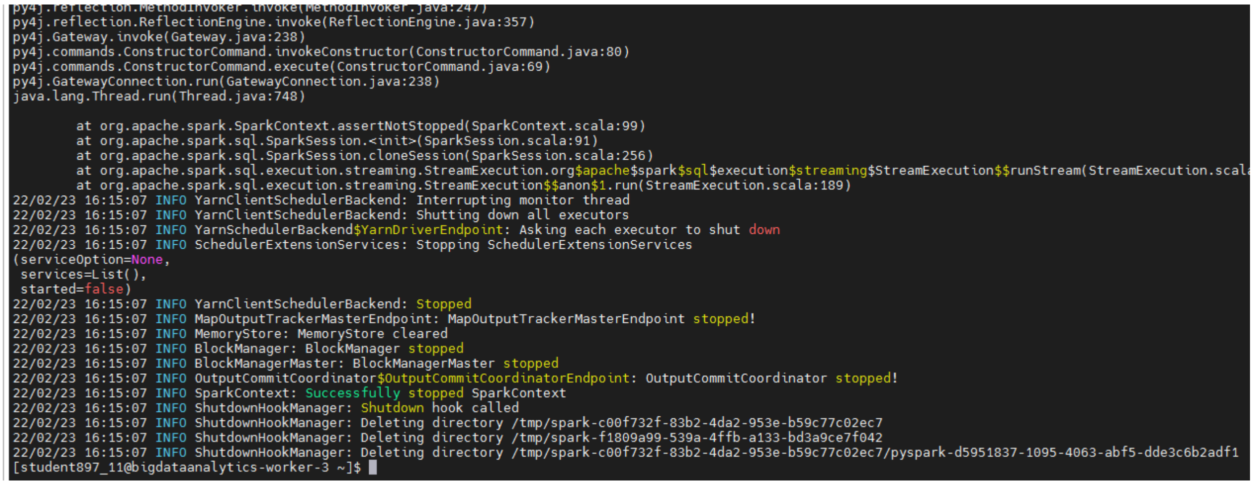
```
[1]+  Stopped                 vi spark-submit-batch.py
[student897_11@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 4 items
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 16:03 .sparkStaging
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 15:53 for_stream
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 15:51 for_stream2
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 16:03 my_submit_parquet_files
[student897_11@bigdataanalytics-worker-3 ~]$
```



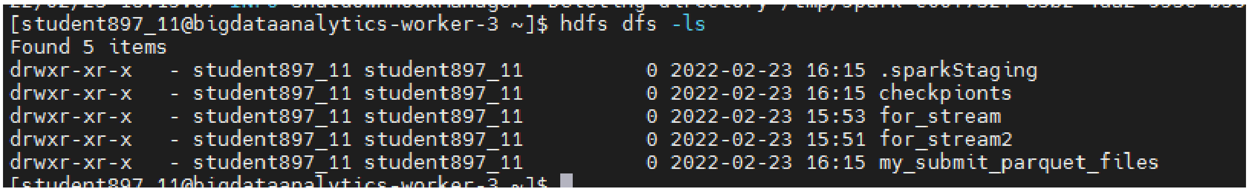
```
[student897_11@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls my_submit_parquet_files
Found 2 items
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 16:01 my_submit_parquet_files/p_date=20220223160153
```



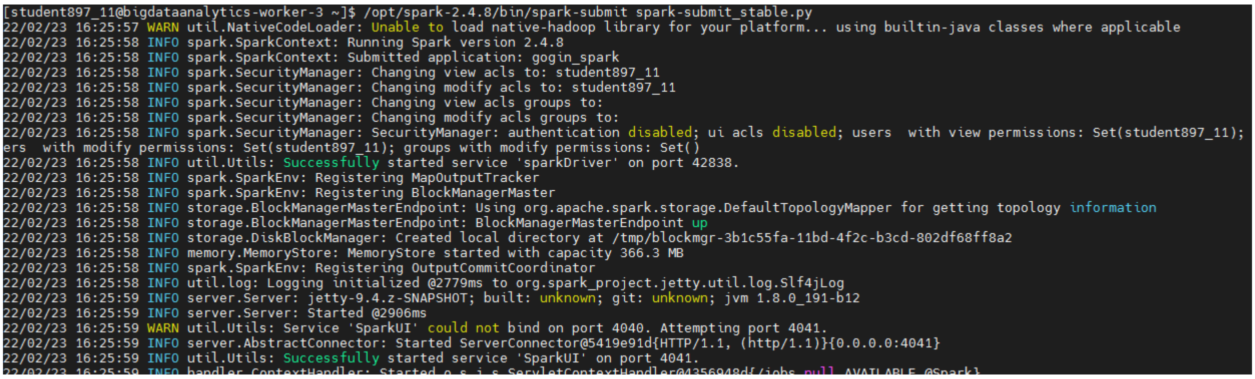
```
[student897_11@bigdataanalytics-worker-3 ~]$ spark-submit spark-submit_stream.py
SPARK_MAJOR_VERSION is set to 2, using Spark2
22/02/23 16:14:51 INFO SparkContext: Running Spark version 2.3.2.3.1.4.0-315
22/02/23 16:14:51 INFO SparkContext: Submitted application: gogin_spark
22/02/23 16:14:51 INFO SecurityManager: Changing view acls to: student897_11
22/02/23 16:14:51 INFO SecurityManager: Changing modify acls to: student897_11
22/02/23 16:14:51 INFO SecurityManager: Changing view acls groups to:
22/02/23 16:14:51 INFO SecurityManager: Changing modify acls groups to:
22/02/23 16:14:51 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions:
ith modify permissions: Set(student897_11); groups with modify permissions: Set()
22/02/23 16:14:51 INFO Utils: Successfully started service 'sparkDriver' on port 46869.
22/02/23 16:14:51 INFO SparkEnv: Registering MapOutputTracker
22/02/23 16:14:51 INFO SparkEnv: Registering BlockManagerMaster
22/02/23 16:14:51 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology infor
22/02/23 16:14:51 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/02/23 16:14:51 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-4810fadc-935a-461b-a307-3fdbdf2718e6
22/02/23 16:14:51 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
22/02/23 16:14:51 INFO SparkEnv: Registering OutputCommitCoordinator
22/02/23 16:14:52 INFO log: Logging initialized @3176ms
22/02/23 16:14:52 INFO Server: jetty-9.3.z-SNAPSHOT, build timestamp: 2018-06-05T17:11:56Z, git hash: 84205aa28f11a4f31f2a3b86d1bb
22/02/23 16:14:52 INFO Server: Started @3288ms
22/02/23 16:14:52 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
22/02/23 16:14:52 INFO AbstractConnector: Started ServerConnector@2d2b148b{HTTP/1.1,[http/1.1]}{0.0.0.0:4041}
22/02/23 16:14:52 INFO Utils: Successfully started service 'SparkUI' on port 4041.
22/02/23 16:14:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@2b435b64{/jobs,null,AVAILABLE,@Spark}
22/02/23 16:14:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@747fe6e8{/jobs/json,null,AVAILABLE,@Spark}
22/02/23 16:14:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@57d1e434{/jobs/job,null,AVAILABLE,@Spark}
22/02/23 16:14:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@76ffdc3f{/jobs/job/json,null,AVAILABLE,@Spark}
22/02/23 16:14:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@7dbc2cae{/stages,null,AVAILABLE,@Spark}
22/02/23 16:14:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@25948769{/stages/json,null,AVAILABLE,@Spark}
```



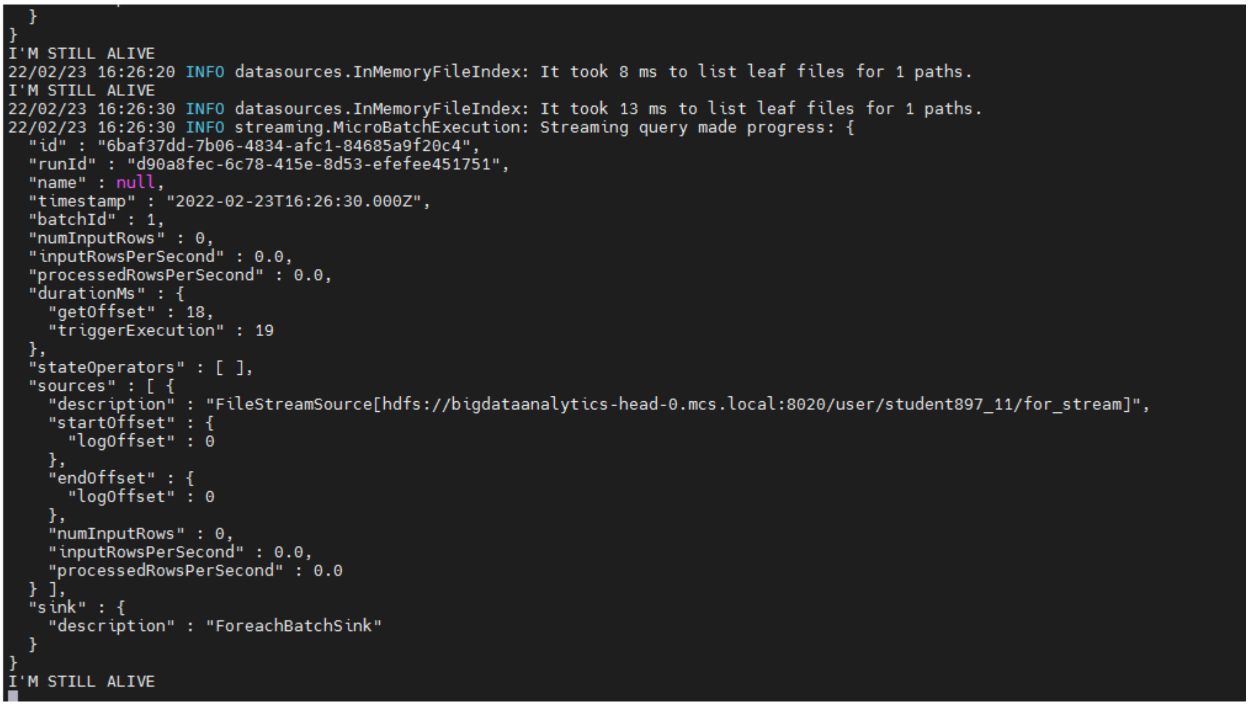
```
py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:247)
py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:357)
py4j.Gateway.invoke(Gateway.java:238)
py4j.commands.ConstructorCommand.invokeConstructor(ConstructorCommand.java:80)
py4j.commands.ConstructorCommand.execute(ConstructorCommand.java:69)
py4j.GatewayConnection.run(GatewayConnection.java:238)
java.lang.Thread.run(Thread.java:748)

        at org.apache.spark.SparkContext.assertNotStopped(SparkContext.scala:99)
        at org.apache.spark.sql.SparkSession.<init>(SparkSession.scala:91)
        at org.apache.spark.sql.SparkSession.cloneSession(SparkSession.scala:256)
        at org.apache.spark.sql.execution.streaming.StreamExecution.org$apache$spark$sql$execution$streaming$StreamExecution$$runStream(StreamExecution.scala
        at org.apache.spark.sql.execution.streaming.StreamExecution$$anon$1.run(StreamExecution.scala:189)
22/02/23 16:15:07 INFO YarnClientSchedulerBackend: Interrupting monitor thread
22/02/23 16:15:07 INFO YarnClientSchedulerBackend: Shutting down all executors
22/02/23 16:15:07 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
22/02/23 16:15:07 INFO SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
22/02/23 16:15:07 INFO YarnClientSchedulerBackend: Stopped
22/02/23 16:15:07 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/02/23 16:15:07 INFO MemoryStore: MemoryStore cleared
22/02/23 16:15:07 INFO BlockManager: BlockManager stopped
22/02/23 16:15:07 INFO BlockManagerMaster: BlockManagerMaster stopped
22/02/23 16:15:07 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/02/23 16:15:07 INFO SparkContext: Successfully stopped SparkContext
22/02/23 16:15:07 INFO ShutdownHookManager: Shutdown hook called
22/02/23 16:15:07 INFO ShutdownHookManager: Deleting directory /tmp/spark-c00f732f-83b2-4da2-953e-b59c77c02ec7
22/02/23 16:15:07 INFO ShutdownHookManager: Deleting directory /tmp/spark-f1809a99-539a-4ffb-a133-bd3a9ce7f042
22/02/23 16:15:07 INFO ShutdownHookManager: Deleting directory /tmp/spark-c00f732f-83b2-4da2-953e-b59c77c02ec7/pyspark-d5951837-1095-4063-abf5-dde3c6b2adf1
[student897_11@bigdataanalytics-worker-3 ~]$
```



```
[student897_11@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 5 items
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 16:15 .sparkStaging
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 16:15 checkpoints
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 15:53 for_stream
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 15:51 for_stream2
drwxr-xr-x   - student897_11 student897_11          0 2022-02-23 16:15 my_submit_parquet_files
[student897_11@bigdataanalytics-worker-3 ~]$
```



```
[student897_11@bigdataanalytics-worker-3 ~]$ /opt/spark-2.4.8/bin/spark-submit spark-submit_stable.py
22/02/23 16:25:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/02/23 16:25:58 INFO spark.SparkContext: Running Spark version 2.4.8
22/02/23 16:25:58 INFO spark.SparkContext: Submitted application: gogin_spark
22/02/23 16:25:58 INFO spark.SecurityManager: Changing view acls to: student897_11
22/02/23 16:25:58 INFO spark.SecurityManager: Changing modify acls to: student897_11
22/02/23 16:25:58 INFO spark.SecurityManager: Changing view acls groups to:
22/02/23 16:25:58 INFO spark.SecurityManager: Changing modify acls groups to:
22/02/23 16:25:58 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(student897_11);
ers  with modify permissions: Set(student897_11); groups with modify permissions: Set()
22/02/23 16:25:58 INFO util.Utils: Successfully started service 'sparkDriver' on port 42838.
22/02/23 16:25:58 INFO spark.SparkEnv: Registering MapOutputTracker
22/02/23 16:25:58 INFO spark.SparkEnv: Registering BlockManagerMaster
22/02/23 16:25:58 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
22/02/23 16:25:58 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/02/23 16:25:58 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-3b1c55fa-11bd-4f2c-b3cd-802df68ff8a2
22/02/23 16:25:58 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MB
22/02/23 16:25:58 INFO spark.SparkEnv: Registering OutputCommitCoordinator
22/02/23 16:25:58 INFO util.log: Logging initialized @2779ms to org.spark_project.jetty.util.log.Slf4jLog
22/02/23 16:25:59 INFO server.Server: jetty-9.4.z-SNAPSHOT; built: unknown; git: unknown; jvm 1.8.0_191-b12
22/02/23 16:25:59 INFO server.Server: Started @2906ms
22/02/23 16:25:59 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
22/02/23 16:25:59 INFO server.AbstractConnector: Started ServerConnector@5419e91d{HTTP/1.1, (http/1.1)}{0.0.0.0:4041}
22/02/23 16:25:59 INFO util.Utils: Successfully started service 'SparkUI' on port 4041.
22/02/23 16:25:59 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4356948d{/jobs,null,AVAILABLE,@Spark}
```



```
  }
}
I'M STILL ALIVE
22/02/23 16:26:20 INFO datasources.InMemoryFileIndex: It took 8 ms to list leaf files for 1 paths.
I'M STILL ALIVE
22/02/23 16:26:30 INFO datasources.InMemoryFileIndex: It took 13 ms to list leaf files for 1 paths.
22/02/23 16:26:30 INFO streaming.MicroBatchExecution: Streaming query made progress: {
  "id" : "6baf37dd-7b06-4834-afc1-84685a9f20c4",
  "runId" : "d90a8fec-6c78-415e-8d53-efefee451751",
  "name" : null,
  "timestamp" : "2022-02-23T16:26:30.000Z",
  "batchId" : 1,
  "numInputRows" : 0,
  "inputRowsPerSecond" : 0.0,
  "processedRowsPerSecond" : 0.0,
  "durationMs" : {
    "getOffset" : 18,
    "triggerExecution" : 19
  },
  "stateOperators" : [ ],
  "sources" : [ {
    "description" : "FileStreamSource[hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student897_11/for_stream]",
    "startOffset" : {
      "logOffset" : 0
    },
    "endOffset" : {
      "logOffset" : 0
    },
    "numInputRows" : 0,
    "inputRowsPerSecond" : 0.0,
    "processedRowsPerSecond" : 0.0
  } ],
  "sink" : {
    "description" : "ForeachBatchSink"
  }
}
I'M STILL ALIVE
```